Abstract

The present invention discloses an apparatus for extracting information from a formatted document, comprising: an input unit (1) for inputting a formatted document; a unit (2) for analyzing the input formatted document and saving the particular typographic information, a unit (3) for identifying special character strings on the basis of the analysis result by means of the typographic information such as font size, character font, color, etc.; a unit (4) for extracting the identified special character strings; and an output unit (5) for outputting the extracted character strings. When the typographic information of a certain character string is determined as a special typographic information, said character string is determined to be special character string. Thus, the present apparatus is able to automatically extract information from different types of format documents.